Probability and Distributions

Probability is a study of uncertainty. It can be thought of as the fraction of times an event occurs, or as a degree of belief about an event.

> Quantifying uncertainty requires the idea of a *random variable*, which is a function that maps outcomes of random experiments to a set of properties that we're interested in.
>
> Associated with the random variable is a function that measures the probability that a particular outcome (or set of outcomes) will occur; this is called the *probability distribution.*

Probability and random variables

- **The sample space Ω**

  The sample space is the set of all possible outcomes of the experiment.

- **The event space $\mathcal{A}$**

  The event space is the space of potential results of the experiment, which is obtained by considering the collection of subsets of Ω, and for discrete probability distributions.

  $\mathcal{A}$ is often the power set of Ω. (Power set include all subsets and empty set

- **The probability P**

  With each event $A \in \mathcal{A}$, associate a number $P(A)$ that measures the probability or degree of belief that the event will occur.

  $P(A)$ is called the probability of A.

The probability of single event must lie in the interval $[0, 1]$, and the total probability over all outcomes in the sample space Ω must be 1, $P(\Omega) = 1$.

Target space $\mathcal{T}$, probabilities on quantities of interest. The elements of $\mathcal{T}$ are referred to *states*.

The function $X : \Omega \to \mathcal{T}$, which takes an element of Ω (an outcome) and returns a particular quantity of interest x, a value in $\mathcal{T}$, is called a *random variable.*

Discrete and continuous probabilities

Discrete probabilities

When the target space is discrete, specify the probability that a random variable X takes a particular value $x \in \mathcal{T}$, denoted as $P(X = x)$.

The probabilities distribution of a discrete random variables is an list of probabilities associated with each of its possible values.

Discrete random variable means the outcome space is discrete.

The expression $P(X = x)$ for a discrete random variable X is known as the probability mass function (`pmf`).

- $\forall a : 0 \leq P(X = a) \leq 1$
- $\sum_a P(X = a) = 1$

Define the *Joint probability* as the entry of both values jointly

$$P(X = x_i, Y = y_i) = \frac{n_{ij}}{N}$$

where n_{ij} is the number of events with state x_i and y_i, and N is the total number of events.

The joint probability is the probability of the intersection of both events, $P(X = x_i, Y = y_i) = P(X = x_i \cap Y = y_i)$.

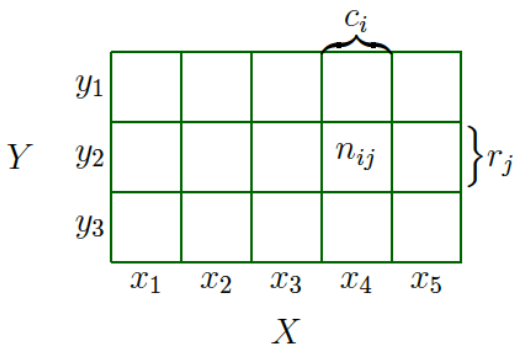


Figure 6.1 Visualization of a discrete bivariate probability mass function, with random variables X and Y. This diagram is adapted from Bishop (2006).

For 2 random variables X, Y, the probability that $X = x$ and $Y = y$ is written as $p(x, y)$ and is called the joint probability.

- The marginal probability that X takes the value x irrespective of the value of random variable Y is written as $p(x)$. $X \sim p(x)$ is the random variable X is distributed according to $p(x)$.

  The marginal probability of each random variable can seen as the sum over a row or column:

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N}$$
$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N}$$

- The conditional probability is when only consider where $X = x$, the fraction of instances for which $Y = y$ is $p(y|x)$.

  The conditional probability is the fraction of a row or column in particular cell:

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$
$$P(X = x_i | Y = y_j) = \frac{n_{ij}}{r_j}$$

Continuous Probabilities

Specify the probability that a random variable X is in an interval, denoted by $P(a \leq X \leq b)$ for $a < b$.

Define. Probability Density Function. A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a *probability density function* (`pdf`) if

1. $\forall x \in \mathbb{R}^D : f(x) \geq 0$
2. Its integral exists and

$$\int_{\mathbb{R}^D} f(x) dx = 1$$

For probability mass function(`pmf`) of discrete random variables, the integral is replace with a sum.

Associate a random variable X with the function f by

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}$ are outcomes of the continuous random variable X. This association is called *law* or *distribution* of the random variable X.

The probability of a continuous random variable X taking a particular value $P(X = x)$ is zero, since

$$P(X = x) = \int_x^x f(x)dx = 0.$$

Define. Cumulative Distribution Function. A cumulative distribution function (`cdf`) of a multivariable real-valued random variable X with states $x \in \mathbb{R}^D$ is:

$$F_X(x) = P(X_1 \leq x_1, \ldots, X_D \leq x_D)$$

where $X = [X_1, \ldots, X_D]^T, x = [x_1, \ldots, x_D]^T$,

The `cdf` can also be expressed as the integral of `pdf` $f(x)$:

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D) \mathrm{d}z_1 \cdots \mathrm{d}z_D \,.$$

Sum, Product Rule, and Bayes' Theorem

Sum rule

which is also known as the *marginalisation property.*

$$p(x) = \begin{cases} \sum_{y \in \mathcal{Y}} p(x, y) & \text{, if } y \text{ is discrete} \\ \int_{\mathcal{Y}} p(x, y) dy & \text{, if } y \text{ is continuous} \end{cases}$$

where $\mathcal{Y}$ are the states of target space of random variable Y.

Product rule

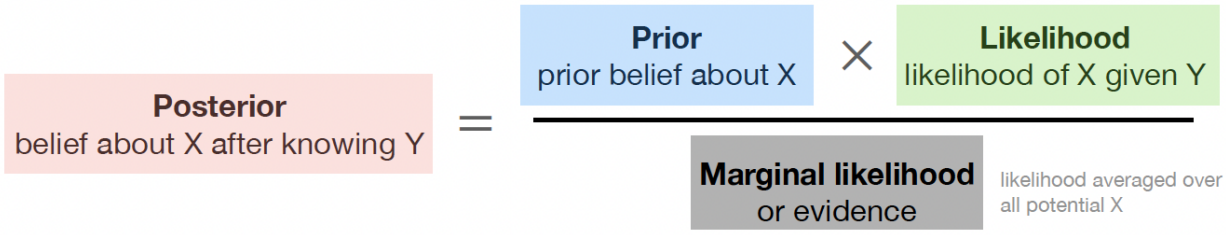
The product rule relates the joint distribution to the conditional distribution by

$$p(x, y) = p(y|x)p(x)$$
$$\text{or } p(x, y) = p(x|y)p(y)$$

This expression is in terms of the `pmf` for *discrete random variable*. For continuous random variables, the product rule is expressed in terms of `pdf`.

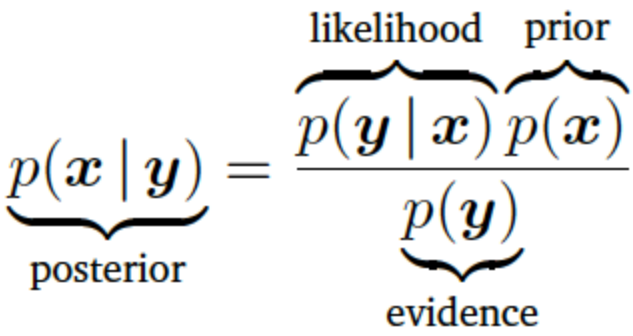
The product rule can be interpreted as the fact that every joint distribution of two random variables can be factorised of 2 other distributions.

Bayes' theorem



Make inferences of unobserved (latent) random variables given that observed other random variables.

With some prior knowledge $p(x)$ about an unobserved random variable x and some relationship $p(y|x)$ between x and a second random variable y, which can be observed. The result about x given y, using the Bayes' theorem:



The *prior* $p(x)$ encapsulates subjective prior knowledge of the unobserved variable x before observing any data. The prior should have nonzero `pdf` or `pmf` on all plausible x.

The *likelihood* $p(y|x)$ is how x and y are related. It's a distribution only in y.

The *posterior* $p(x|y)$ is the quantity of interest in Bayesian statistics, i.e., what we know about x after having observed y.

The *evidence/marginal likelihood* is the quantity:

$$p(y) := \int p(y|x)p(x)dx = \mathbb{E}_X\left[p(y|x)\right]$$

The marginal likelihood integrates the numerator with respect to x, the latent variable. It's independent of x, and it ensures that the posterior $p(x|y)$ is normalised.

Means and covariance

Mean and co-variance are often useful to describe properties of probability distributions (the expected values and spread).

Define. Expected Value. The expected value if a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is:

$$\mathbb{E}_X\left[g(x)\right] = \int_{\mathcal{X}} g(x)p(x)dx$$

Correspondingly, the expected value of a function g of a discrete random variable $X \sim p(x)$ is

$$\mathbb{E}_X\left[g(x)\right] = \sum_{x \in \mathcal{X}} g(x)p(x)$$

where $\mathcal{X}$ is the set of possible outcomes (the target space) of the random variable X.

Consider multivariate random variables X as a finite vector of univariate random variables $[X_1, \ldots, X_D]^T$, the expected values is

$$\mathbb{E}_X[g(\boldsymbol{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D,$$

The expected value is a linear operator. For example, given a real-valued function $f(x) = ag(x) + bh(x)$ where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}^D$,

$$\begin{aligned} \mathbb{E}_X[f(x)] &= \int f(x)g(x)dx \\ &= \int [ag(x) + bh(x)]p(x)dx \\ &= a\int g(x)p(x)dx + b\int h(x)p(x)dx \\ &= a\mathbb{E}_X[g(x)] + b\mathbb{E}_X[h(x)]. \end{aligned}$$

Define. Mean. The mean of a random variable X with states $x \in \mathbb{R}^D$ is an average and is defined as

$$\mathbb{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D, \tag{6.31}$$

where

$$\mathbb{E}_{X_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) \mathrm{d}x_d & \text{if } X \text{ is a continuous random variable} \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) & \text{if } X \text{ is a discrete random variable} \end{cases} \tag{6.32}$$

for $d = 1, \ldots, D$, where the subscript d indicates the corresponding dimension of $\boldsymbol{x}$. The integral and sum are over the states $\mathcal{X}$ of the target space of the random variable X.

In one dimension, there are 2 other intuitive notions of 'average':

- median

  it's the 'middle' value with 50% larger and 50% less than it;

  to continuous values where `cdf` is 0.5

- mode:

  it's the most frequently occurring value

  - for discrete random variable, the mode is the value of x that having the highest frequency of occurrence;
  - for continuous random variable, the mode is a peak in the density $p(x)$

For two random variables, the covariance intuitively represents the notion of how dependent random variable are to one another.

Define. Covariance (Univariate). The covariance between two univariate random variables $X, Y \in \mathbb{R}$ is given by the expected product of their deviations from their respective means

$$\mathrm{Cov}_{X,Y}[x, y] := \mathbb{E}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])]$$

When the random variable associated with the expectation or covariance is clear by its arguments, the subscript is often suppressed.

By using the linearity of expectations, the expression can be rewritten as the expected value of the product minus the product of the expected values:

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

The covariance of a variable with itself $\mathrm{Cov}[x, x]$ is called the *variance* and denoted by $\mathbb{V}_X[x]$.

The square root of the variance is called the *standard deviation* and is often denoted by $\sigma(x)$.

Define. Covariance (Multivariate). If consider 2 multivariate random variables X and Y with states $\boldsymbol{x} \in \mathbb{R}^D$ and $\boldsymbol{y} \in \mathbb{R}^E$, the covariance between X and Y is

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy^T] - \mathbb{E}[x]\mathbb{E}[y]^T = \mathrm{Cov}[y, x]^T \in \mathbb{R}^{D \times E}$$

Define. Variance. The variance of a random variable X with states $\boldsymbol{x} \in \mathbb{R}^D$ and a mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ is

$$\begin{aligned}
\mathbb{V}_X[\boldsymbol{x}] &= \mathrm{Cov}_X[\boldsymbol{x}, \boldsymbol{x}] \\
&= \mathbb{E}_X[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = \mathbb{E}_X[\boldsymbol{x}\boldsymbol{x}^\top] - \mathbb{E}_X[\boldsymbol{x}]\mathbb{E}_X[\boldsymbol{x}]^\top \\
&= \begin{bmatrix}
\mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \ldots & \mathrm{Cov}[x_1, x_D] \\
\mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \ldots & \mathrm{Cov}[x_2, x_D] \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{Cov}[x_D, x_1] & \ldots & \ldots & \mathrm{Cov}[x_D, x_D]
\end{bmatrix}.
\end{aligned}$$

The $D \times D$ matrix is called the *covariance matrix* of the multivariate random variable X. The covariance matrix is symmetric and positive semidefinite.

The diagonal of the covariance matrix are the variances of the *marginals*

$$p(x_i) = \int p(x_1, \ldots, x_D) dx_{\setminus i}$$

where "$\setminus i$" denotes "all variables but i".

The off-diagonal are the *cross-covariance terms* $\mathrm{Cov}[x_i, x_j]$.

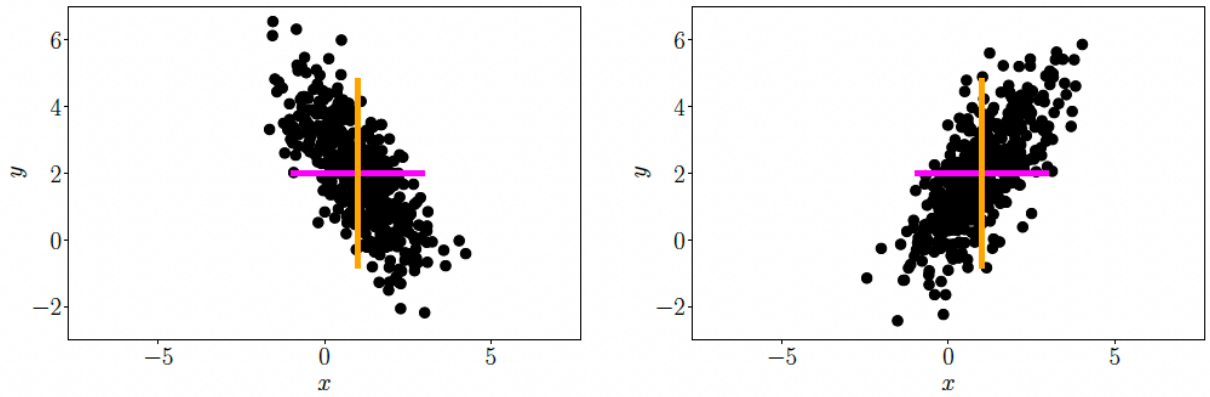
> When compare the covariances between different pairs of random variables, it turns out that the variance of each random variable affects the value of the covariance.
>
> The normalised version of covariance is called the correlation.

Define. Correlation. The correlation between two random variables X, Y is:

$$\mathrm{corr}[x, y] = \frac{\mathrm{Cov}[x, y]}{\sqrt{\mathbb{V}[x]\mathbb{V}[y]}} \in [-1, 1]$$

The correlation matrix is the covariance matrix of standardised random variables, $x/\sigma(x)$. In other words, each random variable is divided by its standard deviation (the square root of the variance) in the correlation matrix.

The covariance (and correlation) indicate how two random variables are related.



(a) x and y are negatively correlated.

(b) x and y are positively correlated.

Positive correlation $\mathrm{corr}[x, y]$ means that when x grows, then y is also expected to grow. Negative correlation means that as x increases, then y decreases.

Empirical Means and Covariances

The above definitions are often called the *population mean and covariance*, as it refers to the true statistics for the population. But in machine learning, we need to *learn from empirical observations of data*.

Two steps from population statistics to realisation of empirical statistics.

- Construct an empirical statistic that is a function of a finite number of identical random variables, $X_1, \ldots, X_N$, with the fact of having a finite dataset (of size N).
- Observe the data the realisation $x_1, \ldots, x_N$ of each of the random variables and apply the empirical statistic.

Given a particular dataset we can obtain an estimate of the mean, which is called *empirical mean* or *sample mean*.

Definition. Empirical Mean and Covariance. The *empirical mean vector* is the arithmetic average of the observations for each variable, and it's

$$\bar{x} := \frac{1}{N} \sum_{n=1}^{N} x_n$$

where $x_n \in \mathbb{R}^D$.

The *empirical covariance* matrix is a $D \times D$ matrix

$$\Sigma := \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})(x_n - \bar{x})^T$$

and empirical covariance matrices are symmetric, positive semidefinite.

Three expressions for the variance

Focus on a single random variable X and use the preceding empirical formulas.

The standard definition of variance

is the expectation of the squared deviation of a random variable X from its expected value μ (the expectation of X),

$$\mathbb{V}_X[x] := \mathbb{E}_X[(x - \mu)^2]$$

The expectation and the mean $\mu = \mathbb{E}_X[(x)]$ are computed depend on whether X is a discrete or continuous random variable.

The variance is the mean of a new random variable $Z := (X - \mu)^2$.

Raw-score formula for variance

$$\mathbb{V}_X[x] = \mathbb{E}_X[x^2] - (\mathbb{E}_X[x])^2$$

which can avoid two passes.

Variance is a **sum of pairwise differences between all pairs of observations**

$$\frac{1}{N^2}\sum_{i,j=1}^{N}(x_i - x_j)^2 = 2[\frac{1}{N}\sum_{i=1}^{N}x_i^2 - (\frac{1}{N}\sum_{i=1}^{N}x_i)^2]$$

It's twice the raw-score expression.

Sums and transformations of Random Variables

Consider two random variables X, Y with states $x, y \in \mathbb{R}^D$, then

$$\begin{aligned}\mathbb{E}[x \pm y] &= \mathbb{E}[x] \pm \mathbb{E}[y] \\ \mathbb{V}[x \pm y] &= \mathbb{V}[x] + \mathbb{V}[y] \pm \mathrm{Cov}[x, y] \pm \mathrm{Cov}[y, x]\end{aligned}$$

Mean and (co)variance exhibit some useful properties when it comes to affine transformation of random variables.

Consider a random variable X with mean μ and covariance matrix Σ and a (deterministic) affine transformation $y = Ax + b$ of x. Then Y is itself a random variable whose mean vector and covariance matrix are

$$\begin{aligned}\mathbb{E}_Y[y] = \mathbb{E}_X[Ax + b] = A\mathbb{E}_X[x] + b = A\mu + b \\ \mathbb{V}_Y[y] = \mathbb{V}_X[Ax + b] = \mathbb{V}_X[Ax] = A\mathbb{V}_X[x]A^T = A\Sigma A^T\end{aligned}$$

Furthermore,

$$
\begin{aligned}
\operatorname{Cov}[\boldsymbol{x}, \boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}(\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b})^\top] - \mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}]^\top \\
&= \mathbb{E}[\boldsymbol{x}]\boldsymbol{b}^\top + \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]\boldsymbol{A}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top\boldsymbol{A}^\top \\
&= \boldsymbol{\mu}\boldsymbol{b}^\top - \boldsymbol{\mu}\boldsymbol{b}^\top + \left(\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\right)\boldsymbol{A}^\top \\
&\overset{(6.38b)}{=} \boldsymbol{\Sigma}\boldsymbol{A}^\top ,
\end{aligned}
$$

where $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ is the covariance of X.

Statistical Independence

Definition. Independence. Two random variables X, Y are *statistically independent* if and only if

$$p(x, y) = p(x)p(y)$$

Intuitively, two random variables X and Y are independent if the value of y (once known) does not add any additional information about x (and vice versa).

If X, Y are (statistically) independent, then

- $p(\boldsymbol{y} \mid \boldsymbol{x}) = p(\boldsymbol{y})$
- $p(\boldsymbol{x} \mid \boldsymbol{y}) = p(\boldsymbol{x})$
- $\mathbb{V}_{X,Y}[\boldsymbol{x} + \boldsymbol{y}] = \mathbb{V}_X[\boldsymbol{x}] + \mathbb{V}_Y[\boldsymbol{y}]$
- $\operatorname{Cov}_{X,Y}[\boldsymbol{x}, \boldsymbol{y}] = \boldsymbol{0}$

The last one may not hold in converse.

i.i.d

In ML, often consider problems that can be modeled as *independent and identically distributed* random variables, $X_1, \ldots, X_N$. For more than 2 random variables,

- independent: refers to mutually independent random variables, where *all subsets are independent*
- identically distributed: all the random variables are from the same distribution

Definition. Conditional Independence. Two random variables X and Y are conditionally independent given Z if and only if

$$p(x, y|z) = p(x|z)p(y|z)\ \forall z \in \mathcal{Z}$$

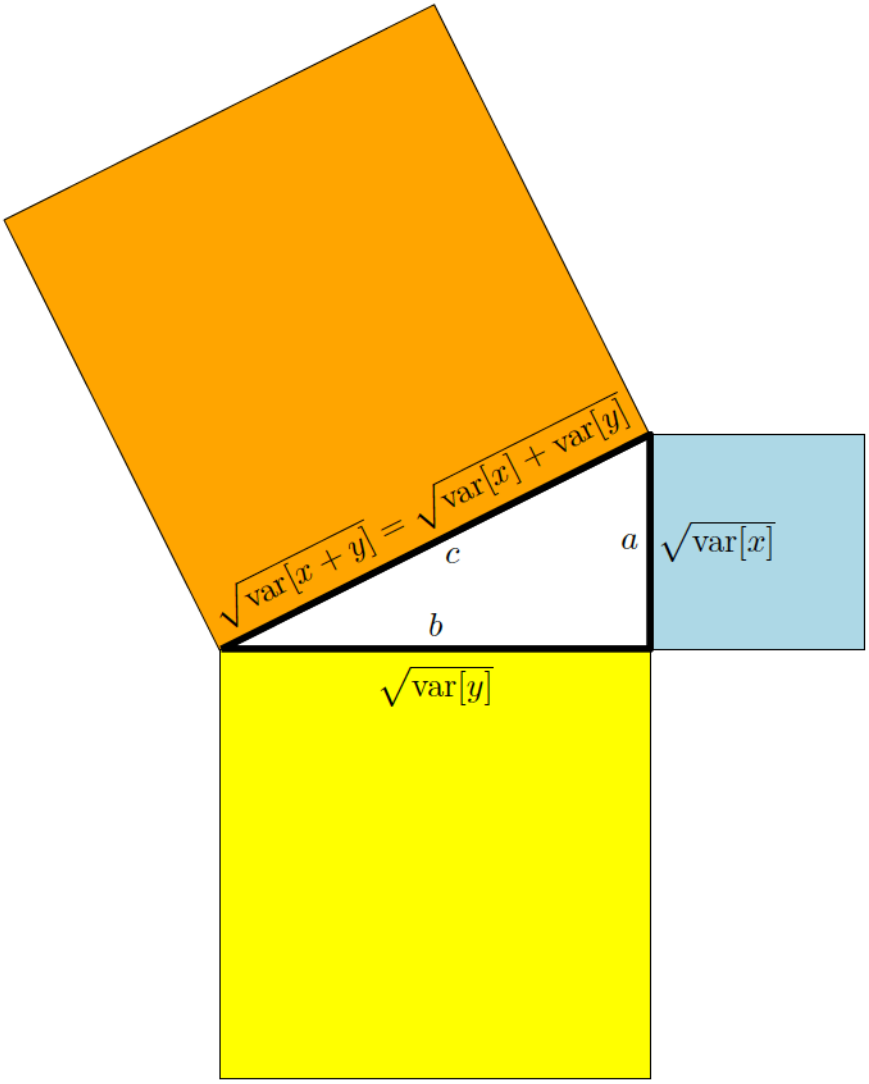
where $\mathcal{Z}$ is the set of states of random variable Z. Use $X \perp\!\!\!\perp Y|Z$ to denote that X is conditionally independent of Y given Z.

Inner products of random variables

Inner product between random variables is, if there are two uncorrelated random variables X, Y , then

$$\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y]$$

The geometric interpretation of the variance relation of uncorrelated random variables.



Gaussian Distribution

The Gaussian distribution is the most well-studied probability distribution for continuous-valued random variables. It's also referred to as the *normal distribution*.

> 🗣️ The Gaussian distribution arises naturally when we consider sums of independent and identically distributed random variables.

For univariate variable, the Gaussian distribution has a density that is given by

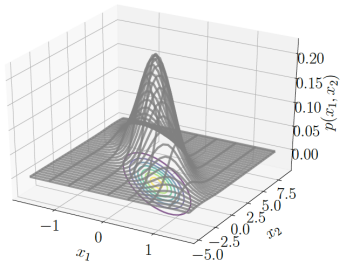
$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

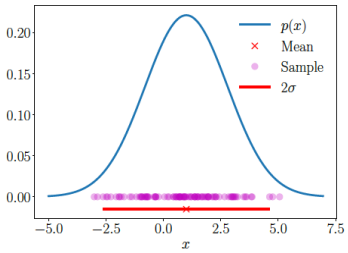
The multivariate Gaussian distribution is fully characterised by a mean vector μ and a covariance matrix Σ and defined as

$$p(x|\mu,\Sigma) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))$$

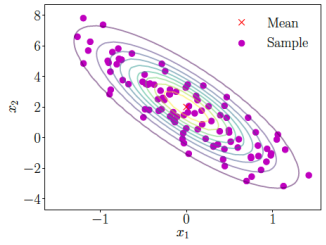
where $x \in \mathbb{R}^D$. Denoted by $p(x) = \mathcal{N}(x|\mu,\Sigma)$ or $X \sim \mathcal{N}(\mu,\Sigma)$.

Bivariate Gaussian(mesh). Univariate Gaussian Bivariate Gaussian





(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.



(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

Special case of the Gaussian with zero mean and identity covariance, $\mu = 0$ and $\Sigma = I$, is referred to as the *standard normal distribution.*

Marginals and conditionals of Gaussians are Gaussians

Let X, Y be two multivariate random variables, that may have different dimensions. To consider the effect of applying the sum rule of probability and the effect of conditioning, explicitly writing the Gaussian distribution in terms of the concatenated states $[X^T, Y^T]$

$$p(x,y) = \mathcal{N}(\begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \begin{bmatrix}\Sigma xx, & \Sigma xy\\ \Sigma yx, & \Sigma yy\end{bmatrix})$$

where $\Sigma_{xx} = \mathrm{Cov}[x,x]$ and $\Sigma_{yy} = \mathrm{Cov}[y,y]$ are marginal covariance matrices of x and y , and $\Sigma_{xy} = \mathrm{Cov}[x,y]$ is the cross-covariance matrix between x and y.

The conditional distribution $p(x|y)$ is also Gaussian and given by

$$
\begin{aligned}
p(x|y) &= \mathcal{N}(\mu_{x|y}, \Sigma_{x|y}) \\
\mu_{x|y} &= \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y) \\
\Sigma_{x|y} &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}
\end{aligned}
$$

In second equation, for computation of the mean, y is an observation.

The marginal distribution $p(x)$ of a joint Gaussian distribution $p(x, y)$ is itself Gaussian and computed by apply the sum rule

$$
p(x) = \int p(x, y)dy = \mathcal{N}(x|\mu_x, \Sigma_{xx})
$$

Product of Gaussian Densities

The product of two Gaussians $\mathcal{N}(x|a, A)\mathcal{N}(x|b, B)$ is a Gaussian distribution scaled by a $c \in \mathbb{R}$, given by $c\mathcal{N}(x|c, C)$ with

$$
\begin{aligned}
C &= (A^{-1} + B^{-1})^{-1} \\
c &= C(A^{-1}a + B^{-1}b) \\
c &= (2\pi)^{-\frac{D}{2}}|A + B|^{-\frac{1}{2}} \exp(-\frac{1}{2}(a - b)^T(A + B)^{-1}(a + b))
\end{aligned}
$$

The constant c can be written as $c = \mathcal{N}(a|b, A + B) = \mathcal{N}(b|a, A + B)$.

> For linear regression (Chapter 9), we need to compute a Gaussian likelihood.
>
> Furthermore, we may wish to assume a Gaussian prior (Section 9.3).
>
> We apply Bayes' Theorem to compute the *posterior*, which results in a multiplication of the likelihood and the *prior*
>
> , that is, the multiplication of two Gaussian densities.

Sums and Linear Transformations

If X, Y are independent Gaussian random variables with $p(x) = \mathcal{N}(x|\mu_x, \Sigma_x)$ and $p(y) = \mathcal{N}(y|\mu_y, \Sigma_y)$, then $x + y$ is also Gaussian distributed and given by

$$p(x + y) = \mathcal{N}(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$$

The mean and covariance matrix can be determined immediately. This is important when considering i.i.d. Gaussian noise acting on random variables.

> Example
>
> Since expectations are linear operations, the weighted sum of independent Gaussian random variables is
>
> $$p(ax + by) = \mathcal{N}(a\mu_x + b\mu_y, a^2\Sigma_x + b^2\Sigma_y)$$

Consider a mixture of two univariate Gaussian densities

$$p(x) = \alpha p_1(x) + (1 - \alpha)p_2(x)$$

where the scalar $0 < \alpha < 1$ is the mixture weights, and $p_1(x)$ and $p_2(x)$ are univariate Gaussian densities with different parameters, i.e. $(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2)$.

The mean of the mixture density $p(x)$ is then given by the weighted sum of the means of each random variable

$$\mathbb{E}[x] = \alpha\mu_1 + (1 - \alpha)\mu_2$$

The variance of the mixture density is

$$\mathbb{V}[x] = [\alpha\sigma_1^2 + (1 - \alpha)\sigma_2^2] + ([\alpha\mu_1^2 + (1 - \alpha)\mu_2^2] - [\alpha\mu_1 + (1 - \alpha)\mu_2]^2)$$

Consider a Gaussian distributed random variable $X \sim \mathcal{N}(\mu, \Sigma)$. For a given matrix A of appropriate shape, let Y be a random variable such that $y = Ax$ is a transformed version of x. So the mean and variance of y are

$$\begin{aligned}\mathbb{E}[y] &= \mathbb{E}[Ax] = A\mathbb{E}[x] = A\mu \\ \mathbb{V}[y] &= \mathbb{V}[Ax] = A\mathbb{V}[x]A^T = A\Sigma A^T\end{aligned}$$

Then the random variable y is distributed to

$$p(y) = \mathcal{N}(y|A\mu, A\Sigma A^T)$$

Conjugacy and the Exponential Family

The class of distributions called the *exponential family* provides the right balance of generality while retaining favorable computation and inference properties.

Bernoulli distribution

The Bernoulli distribution is a distribution for a single binary random variable X with state $x \in \{0, 1\}$. It's governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $X = 1$.

The Bernoulli distribution $\text{Ber}(\mu)$ is defined as

$$\begin{aligned} p(x|\mu) &= \mu^x (1-\mu)^{1-x}, \ x \in \{0, 1\} \\ \mathbb{E}[x] &= \mu, \\ \mathbb{V}[x] &= \mu(1-\mu) \end{aligned}$$

where $\mathbb{E}[x]$ and $\mathbb{V}[x]$ are the mean and the variance of the binary random variable X.

An example of the Bernoulli distribution is when tossing a coin and counting the probability of "heads" up.

Binomial distribution

The Binomial distribution is a generalisation of the Bernoulli distribution to a distribution over integers. In particular, the Binomial can be used to describe the probability of observing m occurrences of $X = 1$ in a set of N samples from a Bernoulli distribution where $p(X = 1) = \mu \in [0, 1]$.

The Binomial distribution $\text{Bin}(N, \mu)$ is defined as

$$\begin{aligned} p(m|N, \mu) &= \binom{N}{m} \mu^m (1-\mu)^{N-m}, \\ \mathbb{E}[m] &= N\mu, \\ \mathbb{V}[m] &= N\mu(1-\mu) \end{aligned}$$

where $\mathbb{E}[x]$ and $\mathbb{V}[m]$ are the mean and the variance of x respectively.

An example where the Binomial could be used is if we want to describe the probability of observing m “heads” in N coin-flip experiments if the probability for observing head in a single experiment is μ.

Beta distribution

The Beta distribution is a distribution over a continuous random variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event.

The Beta distribution $\text{Beta}(\alpha, \beta)$ itself is governed by two parameters $\alpha > 0,\ \beta > 0$ and is defined as

$$p(\mu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

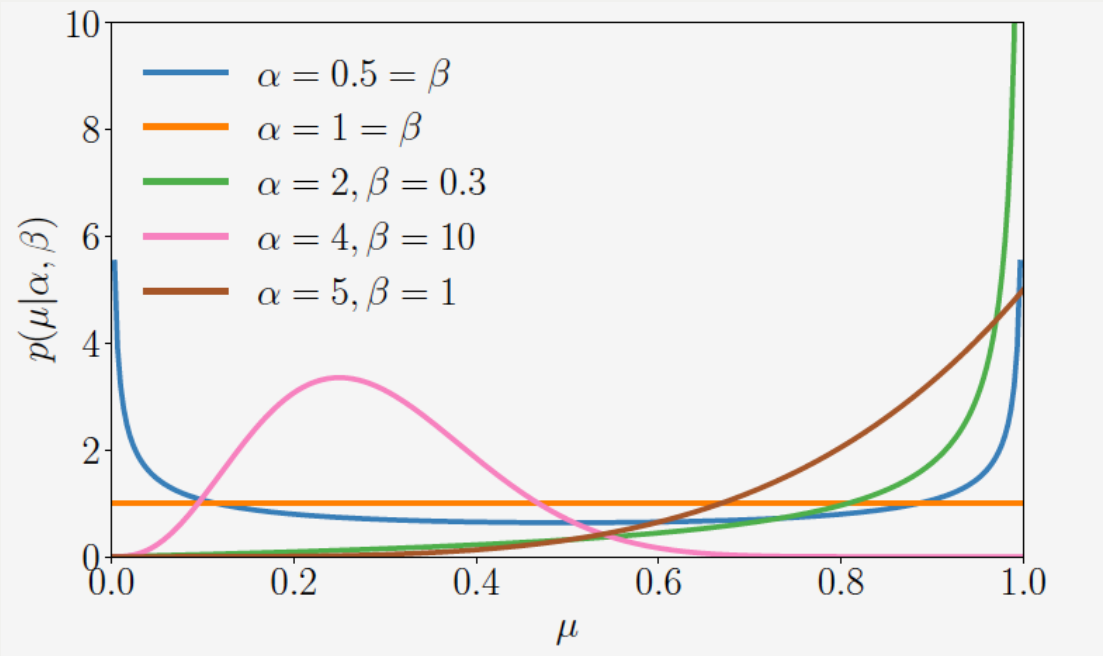
$$\mathbb{E}[\mu] = \frac{\alpha}{\alpha + \beta},\ \mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

where $\Gamma(\cdot)$ is the Gamma function defined as

$$\Gamma(t) := \int_0^\infty x^{t-1}\exp(-x)dx,\ t > 0$$

$$\Gamma(t+1) = t\Gamma(t)$$

Examples of the Beta distribution for different values of α and $beta$



Conjugacy

Computationally convenient priors: conjugacy priors.

Definition. Conjugate Prior. A prior is *conjugate* for the likelihood function if the posterior is of the same form/type as the prior.

> Conjugacy is particularly convenient because we can algebraically calculate our posterior distribution by updating the parameters of the prior distribution.

Examples of conjugate priors for common likelihood functions.

Likelihood	Conjugate prior	Posterior
Bernoulli	Beta	Beta
Binomial	Beta	Beta
Gaussian	Gaussian/inverse Gamma	Gaussian/inverse Gamma
Gaussian	Gaussian/inverse Wishart	Gaussian/inverse Wishart
Multinomial	Dirichlet	Dirichlet

Example 1

Example 6.11 (Beta-Binomial Conjugacy)
Consider a Binomial random variable $x \sim \text{Bin}(N, \mu)$ where

$$p(x \mid N, \mu) = \binom{N}{x} \mu^x (1-\mu)^{N-x}, \quad x = 0, 1, \ldots, N, \tag{6.102}$$

is the probability of finding x times the outcome "heads" in N coin flips, where μ is the probability of a "head". We place a Beta prior on the parameter μ, that is, $\mu \sim \text{Beta}(\alpha, \beta)$, where

$$p(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}. \tag{6.103}$$

If we now observe some outcome $x = h$, that is, we see h heads in N coin flips, we compute the posterior distribution on μ as

$$\begin{aligned} p(\mu \mid x = h, N, \alpha, \beta) &\propto p(x \mid N, \mu) p(\mu \mid \alpha, \beta) && \text{(6.104a)} \\ &\propto \mu^h (1-\mu)^{(N-h)} \mu^{\alpha-1}(1-\mu)^{\beta-1} && \text{(6.104b)} \\ &= \mu^{h+\alpha-1}(1-\mu)^{(N-h)+\beta-1} && \text{(6.104c)} \\ &\propto \text{Beta}(h+\alpha, N-h+\beta), && \text{(6.104d)} \end{aligned}$$

i.e., the posterior distribution is a Beta distribution as the prior, i.e., the Beta prior is conjugate for the parameter μ in the Binomial likelihood function.

Example 2

Example 6.12 (Beta-Bernoulli Conjugacy)
Let $x \in \{0, 1\}$ be distributed according to the Bernoulli distribution with parameter $\theta \in [0, 1]$, that is, $p(x = 1 \mid \theta) = \theta$. This can also be expressed as $p(x \mid \theta) = \theta^x(1 - \theta)^{1-x}$. Let θ be distributed according to a Beta distribution with parameters α, β, that is, $p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1 - \theta)^{\beta-1}$.

Multiplying the Beta and the Bernoulli distributions, we get

$$
\begin{aligned}
p(\theta \mid x, \alpha, \beta) &= p(x \mid \theta)p(\theta \mid \alpha, \beta) && \text{(6.105a)}\\
&\propto \theta^x(1 - \theta)^{1-x}\theta^{\alpha-1}(1 - \theta)^{\beta-1} && \text{(6.105b)}\\
&= \theta^{\alpha+x-1}(1 - \theta)^{\beta+(1-x)-1} && \text{(6.105c)}\\
&\propto p(\theta \mid \alpha + x, \beta + (1 - x)) . && \text{(6.105d)}
\end{aligned}
$$

The last line is the Beta distribution with parameters $(\alpha + x, \beta + (1 - x))$.